BELLSOUTH

EX PARTE OR LATE FILED

Kathleen B. Levitz
Vice President-Federal Regulatory

Suite 900
1133-21st Street, N.W.
Washington, D.C. 20036-3351
202 463-4113
Fax: 202 463-4198
Internet: levitz.kathleen@bsc.bls.com

March 3, 1999

RECEIVED
MAR - 3 1999
FEDERAL COMMUNICATIONS COMMISSION
OFFICE OF THE SECRETARY

EX PARTE

Ms. Magalie Roman Salas
Secretary
Federal Communications Commission
The Portals
445 12th St. SW
Washington, D.C. 20554

Re: CC Docket No. 98-56 and CC Docket No. 98-121

Dear Ms. Salas:

This is to inform you that on March 2, 1999 Ray Lee, Chris Shagnea and I BellSouth, and Dr. Fritz Scheuren and Dr. Mary Batch of Ernst and Young met with members of the Common Carrier Bureau staff. The following Common Carrier Bureau staff members attended at least part of the meeting: Alex Belinfante; Andre Rausch; Florence Setzer; and Daniel Shiman.

During the meeting, BellSouth representatives gave a status report on the workshops that the Louisiana Public Service Commission ("LPSC") staff held on in LPSC Docket No. U22252 – Subdocket C). The purpose of these workshops is to identify the performance measurements, standards and statistical analyses that the LPSC should use to determine whether BellSouth is meeting its statutory obligation to provide CLECs with nondiscriminatory access to UNEs and services. We then focused upon the efforts of Dr. Scheuren working with statisticians at Ernst and Young to develop statistical tests for analyzing performance data to determine whether BellSouth is meeting those statutory obligations. Our subject matter expert used written ex partes made by BellSouth on March 1, 1999 and February 17, 1999 and the attached documents to answer the questions of the Bureau staff about BellSouth's proposed methodology for performing the necessary analysis.

Because the Commission has been considering issues related to performance measurements and standards in both proceedings identified above, we are filing

notice of this ex parte meeting in both dockets, as required by Section 1.1206(b)(2) of the Commission's rules. Please associate this notice with the record of both dockets.

Sincerely,

Kathleen B. Levitz

Kathleen B. Levitz
Vice President – Federal Regulatory

Attachments

cc: Alex Belinfante (w/o attachment)
Andre Rausch (w/o attachment)
Florence Setzer (w/o attachment)
Daniel Shiman (w/o attachment)

FCC Meeting
March 2, 1999

Guided Tour of Filed Documents

February 15 Filing

- Follow-on Statistical Analysis of BellSouth Telecommunications, Inc. Performance Measure Data
- Balancing Type I and Type II Errors
- Analysis of Sprint Data

February 25 Filing

- Follow-on Statistical Analysis of BellSouth Telecommunications, Inc. Performance Measure Data
- "Gaming" the System—Ernst & Young's Response to a Concern about Performance Measurement Testing at an Aggregated Level
- MSA vs. LATA Reporting of Performance Measure Data—Maintenance Average Duration, August 1998

Your Issues

-
-

Time Series

-
-

Next Steps

-
-

Follow-on Statistical Analysis
of
BellSouth Telecommunications, Inc.
Performance Measure Data

The data analysis and data presentation in this report include significant additions and improvements to the Interim Statistical Analysis Report, submitted to the Louisiana Public Service Commission, Docket U-22252, Subdocket C, on November 19, 1998. The changes in the presentation are to provide better documentation and to make the process as nearly self-documenting as possible. In the revised methodology, Ernst & Young has responded to concerns raised at the November workshop, and we have also incorporated additional improvements. The changes in the data analysis are outlined below; a more detailed description of each is then provided. The formula for each calculation is given last section.

Summary of Changes or Additions in the Data Analysis

1. Data Trimming – The FCC has suggested that a "general rule" for trimming the extreme tail of the observations is needed. We have provided one that trims the BST data more severely than in the previous analysis. This rule is used on the Order Completion Interval data.

2. Weighting to the BST Distribution – As requested, we now show the test computed by adjusting or weighting the CLEC observations to the BST distribution, as well as the original analysis which adjusts the BST data to the CLEC distribution.

3. Increasing Sensitivity of the BST Test to Inequality in Standard Deviations – We have made an adjustment to the BST test which will make the test sensitive to unequal variances in the CLEC and BST data, in the same way that the LCUG test is an adjustment to the pooled variance test.

4. Estimate of Variance in the Replicate Test – Because of concerns regarding the choice of variance estimator in the replicate estimate, we now use v_1 as the variance estimator, rather than the more conservative v_2. (Reference: Wolter, K. *Introduction to Variance Estimation,* 1985, Springer Verlag, New York.)

5. Jackknife Test – Because of concerns regarding the replicate technique, we have included an additional test which uses the jackknife approach. This, like the replicate variance estimate, uses the idea of subsample replication and a description can be found in Wolter's 1985 book.

6. When the Data are Uncorrelated – We have added a test of the hypothesis that the adjusted LCUG is suitable for a data set. If this null hypothesis is not rejected, then the adjusted LCUG test procedure can be used. This is done using a two-tailed test of the null hypothesis H_0: "Modified LCUG test statistic" = "Adjusted Jackknife test."

The data provided for the OSS Response Interval does not allow one to use the LCUG modified z test, nor the BST alternatives used on the Order Completion Interval and Maintenance Average Duration data sets. In the Interim Analysis Report we proposed using a modified t test that is based on time series analysis and generalized least squares estimation. This approach is still being used.

Based on the data that we have analyzed so far, Ernst & Young recommends that the Adjusted Jackknife Test described below be used on the aggregated data when the data are reported with enough detail. In cases where the data do not have sufficient detail, alternate approaches like that used for the OSS Response Interval should be used.

Detailed Descriptions

1. Trimming the Extreme Tails of the Distributions

   We have provided a more general trimming rule that trims the BST order completion interval data more severely than in the previous analysis. The completion interval distributions seen up to this point have been skewed, with an extreme tail in only one direction – namely large values. The revised trimming rule in this case is to trim the largest 10 CLEC cases. All BST observations greater than the remaining largest CLEC observation are then trimmed. For example, in the data for the August Order Completion Interval, the 11 largest CLEC observations have the values 24,26,26, 26, 26,27, 28,28,28,34, and 46 days. The 10 observations with values greater than 24 are trimmed from the CLEC data. All BST observations with values greater than 24 are removed from analysis; these trimmed values range in value from 25 to 189 days. This results in 0.22% of the BST data being trimmed and 0.06% of the CLEC data being trimmed.

   Only the Order Completion Interval has been trimmed in this way. The OSS Response Interval data are inappropriate for this type of trimming, and no trimming is needed for percent measurements. The Maintenance Average Duration data has been trimmed at 240 hours in the past, and we have continued to do this. We will investigate applying the new trimming approach on this data in the future.

2. Adjustment by Subclassification to Remove Bias

   Because the data are not the result of a designed experiment but come from an observational study, bias is a serious concern. The true means of the performance measure may differ across classes, defined by time, location, and type of service, and the distribution of the CLEC observations over these classes may differ from the distribution of the BST observations. In this case, under the null hypothesis of no favoritism, the simple difference of means is a biased estimate, and therefore the Type I error is not correct. Adjustment by subclassification is a frequently used devise for trying to reduce such bias. Weighted averages of the subclass

means are compared, using the same weights for the BST cases and for the CLEC cases.

Under the null hypothesis of no favoritism, any definition of the weights, such that the weights add to one, results in an unbiased estimate of the difference. The choice of weights is then made to satisfy other properties of interest. Usually the criteria used for choosing the weights is to minimize the variance of the estimate. The original choice of weights, which adjust the BST observations to the distribution of the CLEC observations, was made because a) it was felt that the distribution of the CLEC's would be the distribution of interest, and b) because we believed that the variance of the estimate using these weights would generally be smaller than the variance of the estimate weighting the CLEC observations to the BST distribution.

Using the same notation as in the Interim Statistical Analysis Report, we have

n_{1j} = the number of BST cases in subclass j

n_1 = the total number of BST cases = $\sum_j n_{1j}$

$\bar{x}_{1j}$ = the mean of the BST cases in subclass j

$\bar{x}_1 = \frac{1}{n_1}\sum_j n_{1j}\bar{x}_{1j}$ = the overall mean of the BST cases

n_{2j} = the number of CLEC cases in subclass j

$\bar{x}_{2j}$ = the mean of the CLEC cases in subclass j

$\bar{x}_2 = \frac{1}{n_2}\sum_j n_{2j}\bar{x}_{2j}$ = the overall mean of the CLEC cases

The estimated difference in the means, adjusted to the CLEC distribution is calculated as

$$\frac{1}{n_2}\sum_j n_{2j}(\bar{x}_{1j} - \bar{x}_{2j})$$

The estimated difference in the means, adjusted to the BST distribution is calculated as

$$\frac{1}{n_1}\sum_j n_{1j}(\bar{x}_{1j} - \bar{x}_{2j})$$

To clarify some apparent misunderstandings, note that if in fact the distribution of the BST's is the same as the distribution of the CLEC's over these subclassifications, then either adjustment results in exactly the same calculation as the simple difference in the means. That is, you still get the correct estimate. In

other words, the adjustment does not in any way "hurt" if in fact it is not needed; in this case, the calculation gives you the simple difference in means.

3. Making the BST test more sensitive to the possibility that the BST variance may be smaller than the CLEC variance.

   The original LCUG test modified the pooled variance test by replacing the pooled variance estimate with the estimate of the BST variance. In a similar manner, an adjustment has been made to the t-statistics calculated using the replicate method and the jackknife method which will increase the absolute size of the test statistic if the estimated BST variance is smaller than the estimated CLEC variance, assuming independence. As with the original LCUG test, the adjusted test statistic will be smaller (less significant) than the unadjusted test statistic when the estimated BST variance is larger than the estimated CLEC variance.

   In general terms, the original BST test statistic is multiplied by the ratio of the estimated standard error of the estimate of the difference (the numerator of the test statistic) under the assumption of independence, divided by the standard error estimate where the CLEC variance estimate is replaced by the BST variance estimate. The exact formula for this adjustment is given in the appendix.

   For the test using the replicate variance estimate, the original statistic for the test is still given on the Decision page and is labeled "REP". The test statistic with this adjustment for disparity in the variances is labeled "REP ADJ."

4. Estimate of Variance in the Replicate Test.

   In the notation of the Interim Statistical Analysis Report, the estimate of variance now used in the calculation of the Replicate t-test is

$$v_1 = \frac{1}{G}\frac{1}{(G-1)}\sum_g (\bar{d}_g - \bar{\bar{d}})^2 .$$

   Reference: Interim Statistical Analysis Report, p. B-8.

5. Jackknife Estimate and Test Statistic

   Another subsample replication technique, called the jackknife, has been included. The jackknife methodology is a broadly useful technique in cases such as this, where the form or the properties of the point estimate are not straightforward. This methodology is used, in general, for two purposes a) to reduce bias, and b) to estimate variance. (Reference: Wolter (1985), *Introduction to Variance Estimation*, Section 4.2.) Using a combination of the notation in Wolter and in the Interim Statistical Analysis Report, the following is a brief description of the jackknife method used here.

An estimator $\hat{D}$ is calculated from the full data set. In the case where the BST observations are adjusted to the CLEC, $\hat{D} = \frac{1}{n_2}\sum_j n_{2j}(\bar{x}_{1j} - \bar{x}_{2j})$ .The observations are then partitioned into G groups. We use the replicates, as defined for the replicate estimate, as the groups for the jackknife test.

Let $\hat{D}_{(g)}$ denote the estimator of the same functional form as $\hat{D}$ but calculated from the observations ***removing*** the gth group. (This is in contrast to the replicate methodology where we calculated the replicate estimate by using only the observations in replicate g.) Then G pseudo-values are defined and used for calculating the mean and variance, where the gth pseudo-value is defined as

$$\hat{D}_g = G * \hat{D} - (G-1) * \hat{D}_{(g)}$$

The estimate of the mean is the mean of the pseudo-values, $\hat{\bar{D}} = \frac{1}{G}\sum_{g=1}^{G}\hat{D}_g$ and the estimate of the variance of $\hat{\bar{D}}$ is $v(\hat{\bar{D}})) = \frac{1}{G(G-1)}\sum_{g=1}^{G}(\hat{D}_g - \hat{\bar{D}})^2$.

The statistic t= $\frac{\hat{\bar{D}}}{\sqrt{v(\hat{\bar{D}})}}$ is distributed approximately as a Student's t with G-1 degrees of freedom. This is the test statistic recorded on the Decision page as the JACK test.

The adjusted jackknife, referred to on the Decision Page as JACK ADJ, is this t-statistic multiplied by the adjustment factor for unequal variances, as described in (3).

6. When the Observations Appear to be Uncorrelated

We found with the data for the performance measure Order Completion Interval that the observations are not independent, but rather there appears to be a clustering effect, or a correlation between observations in the same location. However it appears that the observations for the Maintenance Average Duration, while having different distribution with respect to location may not be correlated. If that is true, then the adjusted or modified LCUG test is appropriate. We have therefore added a test of the hypothesis that the adjusted LCUG test is suitable for the data. If this null hypothesis is not rejected, then the adjusted LCUG test procedure can be used. If the null hypothesis is rejected, then the LCUG test is not appropriate and the BST test should be used.

A two-tailed test of the null hypothesis H_0: "Modified LCUG test statistic" = "Adjusted Jackknife test" is used. (The hypothesis test is made using the

estimates with the BST data adjusted to the CLEC distribution.) This test is performed using a jackknife test. The general jackknife procedure, as described in (5), is applied but now the parameter of interest is not the difference between the BST means and the CLEC means. The parameter of interest is the LCUG test statistic minus the adjusted jackknife test statistic.

Equations

This section provides the equations used for the calculations on the Descriptive Page and the Decision Page of the performance measure analysis reports. The statistical tests used are based on the difference between the mean of the BST and the mean of the CLEC cases. Proportions are means, so these equations also apply to tests based on the difference between proportions or rates.

Notation:

n_1 = the number of BST cases

n_{1j} = the number of BST cases in subclass j

x_{1i} = the value of the performance measure for the i^{th} BST observation

$\bar{x}_1$ = the mean of the BST observations

$\bar{x}_{1j}$ = the mean of the BST observations in subclass j

$$s_1^2 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2}{(n_1 - 1)}$$

Similar notation using the subscript 2 is used to denote the values for the CLEC cases, that is

n_2 = the number of CLEC cases, etc

Adjusted to CLEC

In this case the BST observations are adjusted to the CLEC distribution over the subclasses. The adjusted or weighted mean for the BST cases is

$$\bar{x}_{1w} = \frac{1}{n_2}\sum_j n_{2j}\bar{x}_{1j} = \frac{\sum_j w_{1j}\sum_{i=1}^{n_{1j}} x_{1i}}{\sum_j w_{1j}n_{1j}} \quad \text{where } w_{1j} = \frac{n_{2j}}{n_{1j}} \qquad \text{(E.1)}$$

and the weighted estimate of the BST variance is

$$s_{1w}^2 = \frac{\sum_j w_{1j} \sum_{i=1}^{n_{1j}} (x_{1i} - \bar{x}_{1w})^2}{\sum_j w_{1j} n_{1j} - 1} \qquad \text{(E.2)}$$

The estimate of the difference in means is

$$\bar{x}_{1w} - \bar{x}_2 \qquad \text{(E.3)}$$

and the LCUG test, adjusted to the CLEC's, is

$$\frac{\bar{x}_{1w} - \bar{x}_2}{s_{1w}\sqrt{c_1 + \frac{1}{n_2}}} \quad \text{where } c_1 = \frac{\sum_j w_{1j}^2 n_{1j}}{(\sum_j w_{1j} n_{1j})^2} \qquad \text{(E.4).}$$

The replicate test has been described previously and the jackknife test was described in a previous section. The estimate being calculated in each is the difference in means as in (A.3).

To increase the sensitivity of the BST test to inequality of variances, the jackknife test, and the replicate test, are multiplied by an adjustment factor. Under the assumption that the BST observations are independent and identically distributed (IID) and the CLEC observations are IID, but allowing that the BST and the CLEC observations may have different variances, the expected value of the standard error used in the denominator of the jackknife and replicate tests is

$$\sqrt{c_1\sigma_1^2 + \frac{\sigma_2^2}{n_2}}$$

Therefore to make an adjustment similar to the LCUG adjustment to the pooled variance test, we multiply the jackknife (and replicate test) by

$$\frac{\sqrt{c_1 s_{1w}^2 + \frac{s_2^2}{n_2}}}{s_{1w}\sqrt{c_1 + \frac{1}{n_2}}} \qquad \text{(E.5)}$$

where c_1 is defined in (A.4).

Adjusted to BST

In this case, the CLEC observations are weighted to the distribution of the BST cases. The LCUG test adjusted to the BST is calculated as

$$\frac{\bar{x}_1 - \bar{x}_{2w}}{s_1\sqrt{\frac{1}{n_1} + c_2}}$$

where $c_2 = \dfrac{\sum_j w_{2j}^2 n_{2j}}{(\sum_j w_{2j} n_{2j})^2}$

$w_{2j} = {}^{n_{1j}}\!/_{n_{2j}}$, and

$$\bar{x}_{2w} = \frac{1}{n_1}\sum_j n_{1j}\bar{x}_{2j} = \frac{\sum_j w_{2j}\sum_{i=1}^{n_{2j}} x_{2i}}{\sum_j w_{2j} n_{2j}}.$$

The adjustment factor to the jackknife and replicate test in this case is

$$\frac{\sqrt{\frac{s_1^2}{n_1} + c_2 s_{2w}^2}}{s_1\sqrt{\frac{1}{n_1} + c_2}}.$$

Balancing Type I and Type II Errors

This note is the first of a set of discussions concerning the types of error that are present in hypothesis testing. We first address the issue of balancing the risk of Type I and Type II errors. The important issue of comparing the probability of these errors occurring based on the LCUG modified z test and the alternative test proposed by BellSouth Telecommunications, Inc. (BST) will be addressed at a later date.

Type I error is the error that occurs when the null hypothesis that there is no favoritism on the part of BellSouth is true and we reject it. If we have correctly specified the null distribution, it is controlled directly by the specification of the critical value where we decide to either accept or reject the null hypothesis of no favoritism. Type II error is the error that occurs when the null hypothesis of no favoritism is false but we mistakenly accept anyway. Type II error is not controlled directly but decreases as the sample size increases.

In a controlled experimental study, where the sample sizes are relatively small, it is generally desirable to control the Type I error closely to avoid making a conclusion that there is a difference when, in fact, there is none. The probability of a Type II error is not directly controlled but is determined by the distance between the null hypothesis and the alternative and the sample size. Thus, there is some kind of balance between Type I and Type II errors with Type I error usually controlled more closely.

In Figure 1 below, the distribution assuming the null hypothesis is true is labeled H_0 and the distribution assuming a particular alternative difference between BellSouth and CLEC means is true is labeled H_a. The probability of a Type I error is the area under the null distribution to the left of the test critical value *c*. This region is labeled α. The critical value *c* determines the point beyond which an observed z-value is extreme enough to conclude that BellSouth is favoring itself. This is the decision rule that guides our determination of statistical significance. If, in fact and unknown to us, the alternative distribution is actually the true distribution, we still declare any test statistic that falls to the left of c to be significant. If it falls to the right of *c*, it is not significant. With respect to the alternative distribution, we can see that the area to the left of *c* will lead to an acceptance of the null hypothesis, even though, in this case, it is not true. The probability of a Type II error, incorrectly accepting the null hypothesis for a given correct alternative value, is labeled β on the graphic. Both α and β can be determined for specified null and alternative distributions.

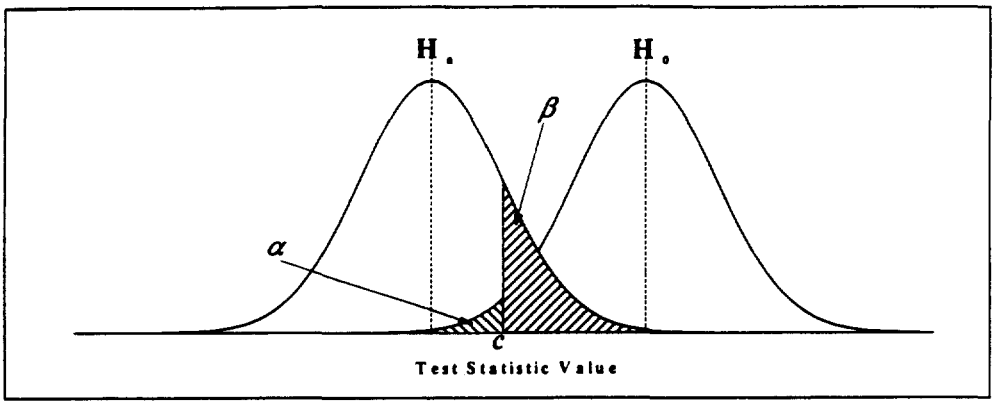

Figure 1

For the purpose of discussion, we consider the modified z test statistic proposed by the LCUG for testing the hypothesis of no favoritism. Let

n_1 = the number of BST observations,

n_2 = the number of CLEC observations,

$\overline{X}_1$ = the average performance measure value of the BST observations,

$\overline{X}_2$ = the average performance measure value of the ~~BST~~ CLEC observations, and

s_1 = the sample standard deviation of the BST observations.

The modified z statistics is

$$z = \frac{\overline{X}_1 - \overline{X}_2}{s_1\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

One interpretation of the null hypothesis that there is no favoritism on the part of BellSouth is that the true means of the BST and CLEC performance measures are equal, as well as the true standard deviations. Suppose that all the observations are independent, and the null hypothesis of no favoritism is true. If the number of BST and CLEC observations are sufficiently large then z has a standard normal distribution. A critical value for the test, given a value for the Type I error α, can be found from a table of the standard normal distribution, or through the use of statistical computer software.

To determine β, we must specifically state the alternative hypothesis. One way to do this is to assume that the true CLEC mean, μ_2, is actually larger than the true BST mean, μ_1, by some fraction of the true BST standard deviation, σ. That is,

$$H_a: \mu_2 - \mu_1 = f\sigma, \qquad f > 0.$$

It can be shown that the probability of a Type II error is given by the area under the standard normal density curve to the right of the value

$$c+\frac{f\sigma}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=c+\frac{f\sigma}{SE_\sigma(n_1,n_2)}. \qquad (1)$$

$SE_\sigma(n_1,n_2)$ denotes the standard error of the mean difference estimator $\overline{X}_1-\overline{X}_2$. The functional notation is used to emphasis the fact that for a fixed value of σ, the standard error varies as the number of observations for BST and CLEC varies.

Figure 2 shows graphs of the probability of a Type II error, β, versus the standard error of the mean difference estimator for $\alpha = 0.05$ ($c = -1.645$) and $f = 0.05$, 0.1, and 0.2. Notice that as the BST sample size, the CLEC sample size, or both sample sizes increase, the probability of a Type II error decreases.

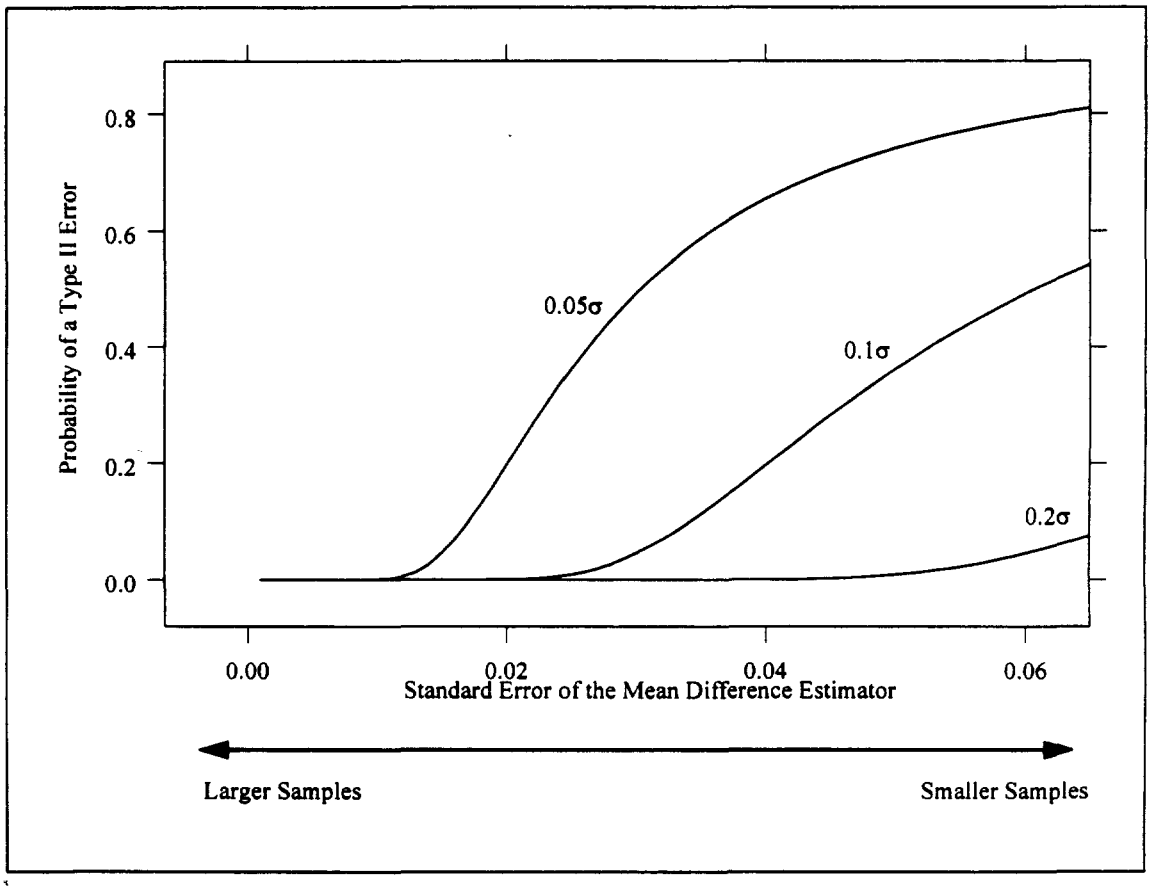

Figure 2: Probability of a Type II Error vs. Standard Error of the Mean Difference Estimator when $\alpha = 0.05$ and the mean difference of the alternative hypothesis is 0.05σ, 0.1σ, or 0.2σ.

In an observational study, where sample sizes are free to vary and may become very large, the balance between Type I and Type II errors can be reversed, with the Type I error risk remaining at a specified level (usually .05 or .01) and the risk of Type II error becoming very tiny. When that happens, we are much more likely to falsely reject a true

null hypothesis of parity than we are to falsely accept an incorrect null hypothesis of parity.

To explore this further, suppose that the number of CLEC observations is some fixed proportion of the number of BST observations, that is, $n_2 = p \cdot n_1$ where $p > 0$. Then (1) can be rewritten as

$$c + \frac{f}{\sqrt{\frac{1}{n_1}\left(1 + \frac{1}{p}\right)}}.$$

Figure 3 shows graphs of the probability of a Type II error, β, versus the Number of BST Observations for $\alpha = 0.05$ ($c = -1.645$), $f = 0.1$, and $p = 0.05$, 0.04, and 0.03. Notice that β drops below 0.05, the value of α, when n_1 is approximately 23,000, 28,000 and 37,000 observations for the respective proportions p.

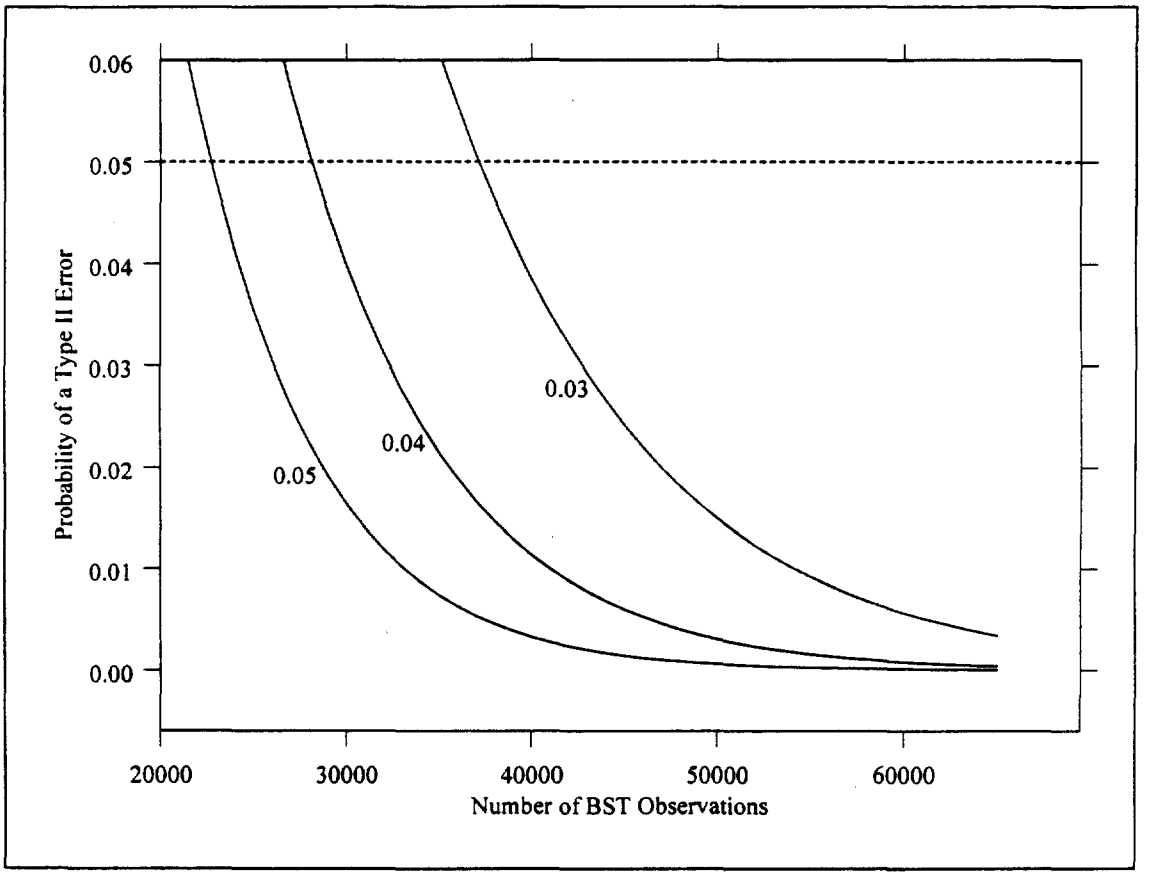

Figure 3: Probability of a Type II Error vs. Number of BST Observations when $\alpha = 0.05$, the mean difference of the alternative hypothesis is 0.1σ and the number of CLEC observations is 0.03, 0.04, or 0.05 times the number of BST observations.

Figure 3 is representative of situations that are possible for the BST/CLEC performance measure data that has been studied. There are many examples where BST has a very large number of observations with the proportion of CLEC observations in the range from

0.04 to 0.05. In these cases, the probability of a Type II error is much smaller than 0.05, the preset probability of a Type I error. To keep a balance between the two types of error, α should be lowered.

There are others issues as well that need to be considered. In an experimental design, the issue of materiality is addressed up front at the design stage in choosing the sample size needed to detect a given difference. This addressing of materiality or business impact often does not occur in the planning stages of an observational study like the BellSouth to CLEC performance comparison. However, it should be addressed in developing the rules that guide a decision of no favoritism or favoritism.

The issue here is not only one of keeping the **risk** of Type I and Type II error in balance; it is, more importantly, an issue of keeping the **costs** of Type I and Type II errors in balance. The cost to BellSouth of spending time and money to pursue the causes of false positives must be balanced against the cost to the CLECs of potential customer loss. Both costs should be explicitly considered. Simulation studies can be done to determine the sample size needed to keep the costs and risks of Type I and Type II errors in balance.

If the result of a statistical test is significant, it should then be compared to a materiality standard to determine whether favoritism exists. If a difference is not statistically significant, even if large enough to exceed the materiality standard, no favoritism exists. In other words, the measured difference must be both accurate enough to trust and large enough to have a business impact.

Analysis of Sprint Data

One of the action items from the November 31 - December 1 Louisiana PSC Workshop was to have Ernst & Young analyze data that Sprint has been using for comparing parity test statistics. On or about January 18th, Ernst & Young received two data sets from Sprint; one file contained 509 CLEC observations and one file contained 21,453 ILEC observations.

On each file, there is a "filing time" and a "resolution time." The variable of interest is the difference between these two times. There is no other relevant information on the files. We have communicated with Dr. Brian Staihr, Regulatory Economist, Sprint, about the data in the files, and he has been very helpful. In a recent email he writes:

> "The sole purpose of that set of data was to provide a means for testing the rejection rates of the modified Z test vs. the standard Z test."

Because of the lack of information on the files, we cannot calculate the proposed BST test nor can we calculate the adjusted LCUG test. Dr. Staihr's email message makes it very clear that these data were constructed for one particular purpose, and they are not necessarily appropriate for any other use.

> "Although the "interval" is a type of performance measurement I want to stress that the purpose of this data set has nothing to do with whether this measurement was correct or not, or whether it is an appropriate measure or not, or whether it even makes sense to use at all! The data is just a bunch of numbers to be used in looking at rejections of null hypotheses and whether or not there are substantial differences in the rejection rates using one Z or another."

In particular, the data cannot be used to evaluate the adjusted LCUG or the BST procedures. As the focus of the discussion has moved past comparing the pooled Z with the modified Z test, there is nothing more to be learned from this data set.

2/15/99

Follow-on Statistical Analysis
of
BellSouth Telecommunications, Inc.
Performance Measure Data

This report includes data analysis and summaries for three additional measures for the five months August 1998 through December 1998. These measures are Percent Missed Installations, Percent Missed Repair Appointments, and Customer Trouble Report Rate.

The findings for Customer Trouble Report Rate are only preliminary. This is the first measure we have considered that is a rate. The numerator is the number of customer troubles reported in an adjustment class (defined by wire center and residence/business). The denominator is the number of lines available in this class. We did not realize in time for this submission that the information on the number of lines available was not disaggregated to the wire center level. This information is now being provided to us.

In addition, while the methodology will not change for this measure, we will need to change the software that calculates the test statistics. The current software assumes the data are in the form of individual observations and the calculations are done as means and sums of squares. We will need to change the programs to work with data for estimating rates. While this is not a difficult conceptual change, it will require that we retest the new programs. Therefore, the full analysis for this measure is not provided here. We have however calculated the original (unadjusted) LCUG tests for the five months and these results are shown below. This should not be considered as a recommendation of the LCUG test statistic for this data.

For the other two performance measures, the data analysis and data presentation in the attachments include the additions and improvements that were described in BellSouth's filing submitted to the Louisiana Public Service Commission on February 19, 1999. As their names imply, both measures are percentages. Because a percentage is calculated as the mean of a variable that takes on only two values, 0 or 1, the, calculations are the same as described in earlier reports.

There is a concern that one should keep in mind when percentages get close to zero or one, and this is discussed below. First, a summary of the changes in the data analysis are outlined. The specific formulas for the statistics were provided in earlier reports and are not repeated here.

Summary of Changes or Additions in the Data Analysis

1. Data Trimming – Trimming is not necessary when the performance measure is a percentage. There are only two possible values of the outcome.

2. Weighting to the BST Distribution – As requested, we now show the test computed by adjusting or weighting the CLEC observations to the BST distribution, as well as the original analysis which adjusts the BST data to the CLEC distribution.

3. Increasing Sensitivity of the BST Test to Inequality in Standard Deviations – We have made an adjustment to the BST test which will make the test sensitive to unequal variances in the CLEC and BST data, in the same way that the LCUG test is an adjustment to the pooled variance test.

4. Estimate of Variance in the Replicate Test – Because of concerns regarding the choice of variance estimator in the replicate estimate, we now use v_1 as the variance estimator, rather than the more conservative v_2. (Reference: Wolter, K. *Introduction to Variance Estimation,* 1985, Springer Verlag, New York.)

5. Jackknife Test – Because of concerns regarding the replicate technique, we have included an additional test which uses the jackknife approach. This, like the replicate variance estimate, uses the idea of subsample replication and a description can be found in Wolter's 1985 book.

6. When the Data are Uncorrelated – We have added a test of the hypothesis that the adjusted LCUG is suitable for a data set. If this null hypothesis is not rejected, then the adjusted LCUG test procedure can be used. This is done using a two-tailed test of the null hypothesis H_0: "Modified LCUG test statistic" = "Adjusted Jackknife test."

Tests of Hypothesis for Percentages when the Percentage is Close to Zero

When testing a hypothesis about a percentage, there are concerns about the applicability of the central limit theorem (the assumption of normality) when the percentage is close to 0 or 1. Several adjustments have been suggested in the literature to improve the test statistic, or the confidence interval, in this case.[1] For the situation here, where we are testing the difference between two proportions, this would be a concern when the percentage of missed appointments, for example, is close to 0, not when the difference between the ILEC percentage and the CLEC percentage is close to zero. In the data presented in this report there are some cases where the estimated percentages are very close to zero. No adjustment has been made to the tests presented here. This is an area that could be considered for improvement in the future.

[1] For example, Newcombe, R. (1998) "Two-sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods," *Statistics Medicine*, 17.

"Gaming" the System
Ernst & Young's Response to a Concern about Performance Measurement Testing at an Aggregated Level

Overview

At the statistical workshop in Baton Rouge, November 1998, the possibility of BellSouth Telecommunications, Inc. (BST) "gaming" the system was discussed. In this context, "gaming" was described as BST giving consistently worse service to a CLEC in a particular location/business area, and canceling this out by giving CLEC's in other locations better than average service. In particular, there appeared to be a feeling that by adjusting the estimates to compare likes to likes, the BST approach in some way made this type of "gaming" possible, while the original LCUG estimate did not.

The original LCUG estimator, applied to disaggregation classes defined in the Service Quality Measurement (SQM) reports, of the difference in service between BST and the CLEC's is the difference between the mean value for the BST and the mean value for the CLEC. This is the numerator of the test statistic. The numerator of the test statistic should be an unbiased estimate of the mean difference under the null hypothesis of parity.

In the Ernst & Young approach, the observational data are divided into classes in order to compare likes to likes. This may mean that one needs to go below the disaggregation levels in the SQM in order to control for confounding factors. These classes are defined in terms of the type of service, the time, and the location of service, so that within a particular class we are comparing "likes-to-likes." The differences are calculated for each class and these differences are aggregated to a total using either the distribution of the CLEC's ("adjusted to the CLEC's") or the distribution of BST. These estimators are referred to as the adjusted estimator. Concern has been expressed that this gives the BST a greater opportunity to "game" the system. We feel this is highly unlikely, as outlined below.

1. *Both the original LCUG estimator and the adjusted estimator use averages or differences in averages and a basic characteristic of averaging is that differences cancel. The unadjusted LCUG has the form of a difference between two means and therefore it also has the characteristic that differences "cancel."*

2. *The fact that sometimes the BST cases have a smaller mean performance measure than the CLEC cases and other times the CLEC cases have a smaller mean performance measure is neither evidence of gaming nor evidence of disparity. One expects some variation. The test statistic requires an estimate of variation in order to test for differences in performance.*

3. *In cases where the identity of the user (BST vs CLEC) is invisible to the provider, gaming would be impossible. In addition, it would generally be difficult and dangerous to attempt to "game" the system in the manner described because the distribution of the observations will not be static over time and the relative*

distribution is not known by BST in advance. This means that the ultimate result of any attempt to distort results is not at all apparent.

Comparing "likes to likes."

In order to compare "likes to likes" it is necessary to compare the same type of service, performed at approximately the same time, in the same geographic or business location. In a designed study, these factors would be controlled in the design of the experiment. In an observational study, these factors must be controlled in the analysis. If there are differences in the means across classes AND if the distributions of the two populations or treatment groups are different across these classes, then the unadjusted difference in the overall means, as used in the LCUG estimator, is a biased estimate under the null hypothesis of parity, and the test does not have the correct Type I error probability.

In the following example, there are only four adjustment classes and the performance measure is the length of time for some type of service. The BST and the CLEC service are in perfect parity; in each class, the mean for the BST cases is equal to the mean for the CLEC cases. But the means differ across classes, and, the distribution across classes differs between BST and the CLEC. For example, class 4 contains 5% of the BST cases and 10% of the CLEC cases.

The unadjusted LCUG estimate is $\bar{x}_1 - \bar{x}_2 = -0.5$ when clearly there is no difference between the two groups and the correct difference is 0. The adjusted estimate is 0 whether adjusting the BST to the CLEC or adjusting the CLEC to the BST.

Example 1. Parity but Different Distributions

Class, j	BST mean x_{1j}	CLEC mean x_{2j}	BST p_{1j} relative distribution	CLEC p_{2j} relative distribution
1	1	1	.30	.20
2	3	3	.30	.30
3	5	5	.35	.40
4	7	7	.05	.10

The unadjusted LCUG estimate is a biased estimate of the difference under the null hypothesis of parity, and therefore the resulting test statistic will not have the correct Type I error probability.

The LCUG estimate also "cancels" differences

If the distribution of the BST cases is the same as the distribution of the CLEC cases across the classes, then the adjusted (BST) estimate results in exactly the same calculation as the original LCUG. This is the case in the following example. There are differences between the BST mean and the CLEC mean within classes, but there is no difference between the distribution of the CLEC's and BST over these classes. In this example it can be seen that there may be differences that "cancel" using the original LCUG estimator too. This is simply a property of averaging.

Example 2. When the Adjusted LCUG equals the Unadjusted LCUG

Class, j	BST mean	CLEC mean	BST p_{1j} relative distribution	CLEC p_{2j} relative distribution
1	2	1	.15	.15
2	4	3	.55	.55
3	5	7	.25	.25
4	7	9	.05	.05

$\bar{x}_1 = \bar{x}_{1w} = 4.1$ and $\bar{x}_2 = \bar{x}_{2w} = 4.0$ and the numerator in the LCUG test would be 0.1.

The adjusted BST estimate is no different from the original LCUG estimate in this respect. Differences can cancel out. And differences do not necessarily imply disparity; random variation is expected. This is why the test requires an estimate of variance.

Another Example

In the following example, the distribution of BST cases over the classes is different than the distribution of the CLEC cases. Therefore, the original, unadjusted LCUG estimate is biased under the assumptions of the null hypothesis. And the example shows that it is not a more robust procedure in terms of "cancellations."

Example 3.

Class, j	BST mean	CLEC mean	BST p_j relative distribution	CLEC p_j relative distribution
1	1	1	.15	.25
2	4	3	.55	.45
3	5	7	.25	.25
4	7	10	.05	.05

The unadjusted means are $\bar{x}_1 = 3.95$ and $\bar{x}_2 = 3.85$ for an estimated difference of 0.1. Adjusting the BST to CLEC gives $\bar{x}_{1w} = 3.65$. Adjusting the CLEC cases to the BST gives $\bar{x}_{2w} = 4.05$. Therefore the estimates of the difference are

	Estimated Difference
Unadjusted LCUG	.1
LCUG Adjusted to CLEC	-.2
LCUG Adjusted to BST	-.1

The Unadjusted LCUG estimate is a biased estimate in this case. The adjusted estimates are both unbiased estimates of the difference between the two groups, if all confounding factors have been accounted for in the definition of the classes.

Again the estimates of the difference are not sufficient for testing for a difference between the two groups. A variance estimate is needed. Adjusting to the CLEC is preferred

because, in general, we expect that the variance of the LCUG adjusted to the CLEC data will be smaller than the variance when adjusting to the BST data.

Sensitivity to the Distribution of the CLECs

Finally, while highly unlikely, even if it were possible to "game" the system as proposed, it would be nearly impossible to predict the outcome because of the great number of classes and because of the possible changes in the volume of orders, for CLEC orders in particular. The data are coming from an observational study and there is no control over how many CLEC orders in a particular location, or with a specific type of service, may occur one month. It is not known in advance what the volume of the service will be over the adjustment classes, that is, over the types of service by location and time. Therefore even if it were physically possible to identify which customers were CLEC vs BST customers, and if it were possible to completely control the level of service, the type of "gaming" that has been described would be very difficult and dangerous to attempt.

Take the much simplified example that has been used here, with only four adjustment classes, compared to the 100's of possible adjustment classes in the real situation. Assume that the true average for this performance measure is as shown in Example 1. And suppose that in the past, the distribution of the BST cases is that shown in Column 4, and the distribution of the CLEC cases in the past is given in column 5 in Example 4 below. Suppose further that BST is capable of controlling service so that the mean service could be controlled to turn out as shown in columns 2 and 3, where the BST cases receive worse (longer) service in class 2 but better (shorter) service in classes 3 and 4.

Example 4.

1. Class	2. BST mean	3. CLEC mean	4. BST p_{1j}	5. "Prior" CLEC p_{2j}	6. "New" CLEC p_{2j}
1	1	1	.15	.20	.27
2	4	3	.55	.70	.45
3	5	7	.25	.05	.20
4	10	12	.05	.05	.08

With this scenario, if the CLEC distribution remains as shown in column 5, then we would have the following estimates of the difference:

	Estimated Difference with CLEC Column 5
Unadjusted LCUG	.85
Adjusted to CLEC	.50
Adjusted to BST	-.05

Then even if the standard error of the estimated difference is very small, for example if the standard error is .05, there is no evidence of a significant difference in the mean performance.

But the distribution of the types of orders or types of cases is not under control, and in particular the distribution of the CLEC's will not necessarily be static over time. Suppose that the distribution of the CLEC's changes, to the distribution shown in column 6. Then the following are the estimates of the difference:

	Estimated Difference Using Column 6
Unadjusted LCUG	.12
Adjusted to CLEC	-.11
Adjusted to BST	-.05

In this case, if the standard error is small, say .05, then the test using the estimate adjusted to the CLEC will result in a statistically significant difference between the BST mean and the CLEC mean.

Thus, even if the BST were able to identify CLEC services at the point of delivery and in some way provide diminished service to the CLEC, the ultimate outcome of such efforts is not at all clear in terms of its effect on significance tests.

Summary

The original LCUG estimator is not applicable in this situation with observational data because it is not generally unbiased under the assumptions of the null hypothesis.

Both the original LCUG estimator and the BST adjusted estimator are calculated as differences in averages. Such estimators have the property that differences cancel. The BST adjusted estimator is not more or less sensitive in this way than the original LCUG estimator.

One cannot make claims about differences in treatment by looking only at the estimate of the difference in means. In order to test the hypothesis, an estimate of the variability is required.

In cases where the identity of the customer is invisible to the provider, gaming would be impossible.

Even if it were physically possible to "game" the system as described, because the volume of the cases is not controlled, it would be extremely difficult to attempt, and, we believe, unlikely to succeed.

MSA vs. LATA Reporting of Performance Measure Data
Maintenance Average Duration
August 1998

The attached tables and charts show the results of disaggregating the August 1998 Non-designed Maintenance Average Duration performance measure data to both the Metropolitan Statistical Area (MSA) level, and the Local Access Transport Areas (LATA) level. The displays are presented in the same style as with other performance measure data: a descriptive page showing basic descriptive statistics, and a decision page showing statistical test results.

Note that the replicate and jackknife test statistics exhibit unstable characteristics in some of the MSAs. This is due to the fact that LATA were taken into account when the replicates were constructed, but MSA were not. More work needs to be done in order to adjust these statistics when considering MSA. Methodology does exist for this, but it has not been applied because, for the August data, we have concluded that the adjusted LCUG modified z test statistic is statistically equivalent to the adjusted jackknife test statistic. The table below summarizes the results.

LATA	Test Statistics (Adjusted to CLEC Adjusted to BST)	MSA	Test Statistics (Adjusted to CLEC Adjusted to BST)
Shreveport	-1.47 -1.06	Alexandria	-0.42 -0.64
		Monroe	0.05 -0.43
		Shreveport	-0.85 -0.35
Lafayette	0.18 -0.80	Lafayette	0.71 0.55
		Lake Charles	-0.46 -1.28
Baton Rouge	-0.27 -0.57	Baton Rouge	-0.50 -0.76
New Orleans	-1.65 -1.04	Houma-Thiboaux	0.48 0.78
		New Orleans	-1.94 -1.43
Louisiana State	-1.88 -1.69	Outside MSA	-0.89 -0.82

Based on these data, Ernst & Young is not sure that any extra insight is gained in comparing LATA and MSA. It is Ernst & Young's understanding that BellSouth does

not have a preference between these choices for substate reporting. However, BellSouth does not want to base testing decisions on data below the state level.

The original decision to report substate statistics at the LATA level was made by Ernst & Young for the following reasons:

- LATAs are a meaningful geographic business unit for BellSouth. MSAs are statistical entities, subject to redefinition by the Office of Management and Budget (OMB) of the Federal government. In fact, there will be a major revision of MSAs in connection with the 2000 Census. Hence MSA units may not have a stable definition over time.

- MSA's in Louisiana vary considerably in size as measured by the number of wire centers servicing them. For several MSAs, this has the effect of making the available sample sizes "small" – too small to safely employ the types of statistical test that are used without modification.

It was suggested at the November 30, 1998 Louisiana Public Service Commission Workshop that one reason for using MSAs as the substate reporting level is that this provides an urban vs. rural comparison. It is well documented in the statistical literature that MSAs do not provide good urban/rural comparisons.[1] Therefore, Ernst & Young's opinion on this matter is that a telecommunications business unit below the state level, such as LATA or TURF, is a more appropriate substate geographic reporting level.

[1] For a discussion of rural vs. urban measures see, Goodall, C. R., Kafadar, K., and Tukey, J. W. (1998) "Computing and Using Rural versus Urban Measures in Statistical Application," The American Statistician, 52, 101-111.

Response to Dr. Colin Mallows' Comments Originally Read at the Louisiana Public Service Commission Workshop on November 30, 1998.

Dr. Mallows' made a series of detailed comments at the November Louisiana Public Service Commission (LPSC) Workshop. These have been reproduced below, followed by our response. Before going into the specifics, some general observations may be worth making. The most important of these is that we are very appreciative of Dr. Mallows' ideas. They have led both to improvements in our thinking, and, we hope, to its exposition.

We do not necessarily agree with all of his comments, but we believe that our differences can be summed up by a statement Dr. Mallows made in the American Statistical Association's 1997 Fisher Memorial Lecture.

> "In a complex problem, it is possible for ethical analysts to take opposing positions. But this style of thinking is what statisticians should be trained to do."

Dr. Mallows' views (in *italics*) appear as they do in a document forwarded to us by Jay Bradbury of AT&T. We have, however, broken down his statements numbered 1 through 8, into substatements in order to clarify exactly what we are responding to.

Statement No. 1

1.1 *The BST team has done a good job of descriptive data analysis. They have made many sound comments on the importance of data-verification, the need to trim outliers, the importance of disaggregation, and the need to identify confounding variables and to adjust for their effects.*

We thank Dr. Mallows for his compliments. The Ernst and Young approach was to recognize that the data are an example of an observational study and the resulting methodology is based on the associated literature. An observational study uses data that come from a process where there was neither a design nor a random assignment of treatments.

1.2 *They have made a useful contribution by showing how the BST data can be adjusted to make it directly comparable to the CLEC data*

It is imperative that adjustments are made in order to compare "likes-to-likes." Bias is the primary concern in observational studies. In order to compare BST and CLEC data, it is necessary to consider any variables that are known or suspected to have an important relationship with the performance measure. In the design of such a study, variables accounting for time and location are generally considered. Therefore we recommend using a location category (wire center) and a time variable in the comparison for any performance measure.

The adjustment we employ is commonly used in observational studies when there is a considerable amount of data involved. To our knowledge it was introduced in Cochran (1968) "Removing Bias in Observational Studies," *Biometrics*. Thus, we should refer to it as the Cochran adjustment.

1.3 *However I think their conclusions are not supported by the evidence that they have presented. They have not shown that the FCC/LCUG approach is invalid.*

Our conclusions are about the data that we have analyzed, and not necessarily about the general validity of the LCUG/Pooled variance[1] approach. We have concerns when the data exhibit a dependence structure. We have never seen any discussion of this notion in LCUG documents.

We assumed that the LCUG approach was targeted at the data presented in the Service Quality Measurements (SQM) reports, and we find both the LCUG and Pooled tests inadequate for use on these data. Straightforward use of these tests can result in biased estimates of the difference in means, incorrect variances, and hence, inappropriate test statistics.

It should be noted that our methodology, and the LCUG/Pooled approaches are all basically equivalent when there is no dependence structure in the data. This point is discussed in **Technical Appendix A.1**.

1.4 *At one point they have made an adjustment that favors BST, is in the wrong direction, and may be quite large.*

We believe this statement refers to our choice of variance estimator that Dr. Mallows discusses in his fifth point. We chose the method that was recommended as conservative in Wolter (1985) *Introduction to Variance Estimation*, Springer-Verlag. However, we now realize that this definition of conservative does not coincide with what LCUG feels is conservative. Namely, one should always err on the side of the CLECs.

As is shown below in discussing point No. 5 (specifically 5.4), there is very little difference in the two variances. However, we understand LCUG's concerns, and will use the smaller of the two variances in future computations.

[1] The pooled variance Z-test was mentioned in footnote 1 in the FCC's **Notice of Proposed Rulemaking (Appendix B)** This has become known as the FCC approach, however, the FCC does not approve or disapprove of its use. We will therefore more appropriately now refer to it as just the "pooled" approach.

Statement No. 2

2.1 *On page 41 the BST analysts remark that for the Average OSS Response Interval they only had daily summary averages to work with, and that this sever[e]ly limited their approach to analyzing statistical significance. Clearly if BST does not make suitable data available, any statistical approach will be handicapped.*

The word "severely" applies to our ability to use the LCUG/Pooled approaches on the data. Since we could use a time series approach to analyze and test the data, there is no need to have the data in another form. Just because one cannot use a particular tool on data does not mean that there is something wrong with the way the data are structured.

2.2 *The BST analysts seem to have had access to the numbers of orders, since they used these to adjust the BST data; but they have not presented these numbers in this report.*

There are disclosure issues involved when releasing data, and we must be sensitive to this issue. For this reason, the specific BST and aggregate CLEC counts were not provided. We did omit the OSS Response Interval SQM in Appendix G. This was not intended, and we will add it when the report is updated.

Statement No. 3

3.1 *The BST analysts claim in their summary Table 1 that their recommended methods will have essentially the same power as the FCC and LCUG tests to detect differences, should they exist.*

The term "power" should not have been used in this context. The point we were trying to make referred to efficiency of the test (as required by the LPSC order we were addressing). The word "efficiency" we are interpreting as confidence interval length.

3.2 *They give no evidence of this, and in fact in many of their summary tables the BST statistic is less extreme than are the FCC and LCUG statistics, which suggests that it has less power.*

We agree that we have not given specific evidence. Any discussion of a comparison of power needs to start with defining a specific alternative hypothesis that would be considered a significant degradation in services for the CLECs. One example given by LCUG at the workshop involves studying a test statistics behavior when a difference in the means is equal to 10 percent of the BST standard error. We will use this example in one of the follow-ups requested by the LPSC from the workshop.

We point out, however, that the LCUG/Pooled estimator applied to the original SQM data is not a fair measure because the estimate of the mean difference (the numerator in the LCUG/Pooled test statistic) is biased. This can be corrected by using the adjusted difference in the numerator, as we have done. However, when there is dependence between observations, the estimate of the variance (in the denominator) is also incorrect. Therefore, one cannot infer from looking at the test statistics alone anything as regards differences in power.

3.3 *On page B-13 they claim "there is a minimal loss of power using the replicate method compared to the FCC or LCUG method (2.04 vs. 2.00 for the 5% two-sided significance level)". But here they are only comparing the critical values of the tests, and this says nothing about the powers of the tests.*

We agree that it is incorrect to use the term power in this context. Due to the problems with the LCUG and Pooled tests when applied to dependent data, we chose to compare critical values of the test. It is appropriate to say, see our answer in 3.1 above, that there is very little loss of "efficiency."

If data are independent, and the replicate variance estimate is adjusted so that it is sensitive to differences in variance, then it can be shown that the results of the LCUG and BST tests are similar. In this situation then there will be little loss of power using the BST test. (See **Technical Appendix A.1-2.**)

Statement No. 4

4.1 *The analysts assert that the LCUG and FCC procedures require strong assumptions that are not warranted in the data they have examined.*

As we have stated previously, we assumed that the LCUG/Pooled methodology was to be applied to the data at the levels of disaggregation reported in the SQM. In order for these methods to be applied, one must assume that the observations are independent and identically distributed.

The exploratory analysis that we performed on the data sets indicated that this was too strong an assumption to make. Thus, we did not feel that these procedures should be used unchanged.

It should be noted, however, that our findings do not imply that we must test within each possible adjustment class. This is neither practical nor necessary. The methodology proposed by Ernst and Young results in an estimate of the difference aggregated over all groups. This would be unbiased if all the variables that can affect performance have been accounted for in the classes. At the very least, it would result in tests at the present level of aggregation that have less bias then the proposed LCUG test.

4.2 *These procedures have three components. First, a particular statistic is chosen. This is some function of the BST and CLEC data, designed to be sensitive to the kinds of violation of parity that are deemed to be important. The FCC proposed a standard form of the two-sample t statistic; LCUG proposed a modification of this. The BST analysts rely on the difference between the CLEC mean and an adjusted BST mean.*

We do not agree. The replicate methodology employed is also a modification of the standard form of the two-sample "t" statistic. As Dr. Mallows points out in remark No. 1.2, the Cochran adjustment used on the data is necessary in order to make the BST data directly comparable to the CLEC data.

4.3 *It is the judgement of LCUG that a simple comparison of means will not be responsive to all of the possible ways parity might be violated.*

We agree with this point. The LCUG test statistic is a variation on the standard pooled variance test of the difference between two sample means. It has been modified to be sensitive to certain differences in variances as well. While the original method we proposed lacks this sensitivity, a simple adjustment can be done to our test to give it a similar property. We discuss this in **Technical Appendix A.2**.

We note, though, that this test methodology is not the same as testing whether the distribution of the aggregate CLEC values is the same as the distribution for the BST values for a particular performance measure. Testing for equality of distributions is a more complicated problem.

4.4 *It is easy to provide scenarios in which parity is being violated but a comparison of means shows no effect. BST has not presented evidence that the only differences that occur are shifts in means, with variances staying the same.*

We do not argue the point that scenarios can be constructed in which parity is being violated but a comparison of means shows no effect. However, in the data we examined, more often than not, the CLEC variance was smaller than the BST variance. This being the case, the variance sensitivity adjustment makes the test less likely to detect instances where BST is favoring itself in terms of a difference between the means.

4.5 *The choice of statistic does not depend on any assumptions; though of course the efficacy of the resulting procedures will depend on how the data actually behave.*

As we have stated, the data exhibit a wire center dependency which precludes the use of the LCUG or Pooled procedure at the levels of disaggregation reported in the SQM.

It is our understanding that LCUG wants to handle this through deeper disaggregation, and testing at this very deep level. We do not believe that this removes the dependency problem since data from the same wire center is still dependent, despite being disaggregated.

Even if the dependency problem is ignored, the deep disaggregation will most likely call for testing procedures that are suitable when sample sizes are small. LCUG suggests using a permutation test for this situation (letting the computer draw many pseudo-random samples).

This is not practical and it is not necessary. We have presented a way to avoid costly testing at very deep levels of disaggregation. Dr. Mallows agrees that BST data can be adjusted to make it directly comparable to the CLEC data, so we can use it at a high level of aggregation.

In Dr. John Jackson's recent submission to the LPSC, "Using Permutation Tests to Evaluate the Significance of CLEC vs. ILEC Service Quality Differentials," he notes that permutation tests he ran were taking 15 to 20 minutes to complete. Even with an improved algorithm and a faster computer, these tests might take five seconds on average to complete.

In the case of just one performance measure, "Order Completion Interval," this could necessitate possibly 16,000 tests. If this had to be done for all performance measures, at very deep levels of disaggregation, the number of tests could easily reach 100,000. Thus, it could take 500,000 seconds, or approximately six straight days for the computer to just perform the tests on the Louisiana data. If this had to be done in all nine states that BellSouth operates in, it would take nearly two computer months to process the test results for just one calendar month of data.

4.6 *The second component of the FCC and LCUG procedures concerns the choice between a one-sided and two-sided test, and the size of the test (the type 1 error). Since the objective of the analysis is to check whether the CLECs are being given service that is at least equal in quality to what BST provides itself, it seems to LCUG that one-sided tests are appropriate.*

We disagree. In instances where it appears that BST is favoring itself, action needs to be taken to correct the problem. This does not mean, however, that there is no information of value when it is learned that BST may be favoring the CLECs.

4.7 *I do not dispute that both BST and the CLECs may be very interested to find that in some cases the CLEC is getting better service that BST, but for the purpose of checking compliance this is irrelevant.*

Again, we disagree. When looking at the results of tests over time, or even at the results of tests at different levels of disaggregation, it is important to know if there

are significant results in both direction. This can provide an indication of whether significant results are random occurrences, or a systematic problem. It also provides information on the stability of the process.

4.8 *As for the choice of type 1 error, in the BST analyses the conventional level of 5% two-sided, equivalent to 2 1/2% one-sided, is used. LCUG has argued that the (one-sided) type 1 error should be rather larger than this, since while this small value does protect BST from being falsely accused when it is in compliance, it necessarily implies a large probability that a truly important violation, if it occurs, will fail to be detected.*

LCUG has , in fact, offered different one-sided levels of significance at different times as their filing in Louisiana makes clear. It is true that the larger the (one-sided) Type I level of significance is set, *ceteris paribus*, the smaller will be the Type II error. Choosing the right balance here is a hard problem. Even so, it is not necessarily true that there exists "a **large** probability that a **truly important** violation will fail to be detected." (emphasis added) As noted elsewhere (see No. 3.2 above), we will be looking at this issue directly for the Commission.

This might be a place to add in a reminder of something that we said over and over at the workshop. **It is very difficult to use observational studies to show causality or in this case disparate treatment.** Therefore, even if we find a difference between BST and the CLECs on a measure, it is not necessarily proof of disparate treatment. This is why a "drill down" is needed to investigate the cause of the differences. These may be differences due to factors that affect the performance measures that were not included in the Cochran adjusted estimate, or the differences may be due to disparate treatment. But this cannot be determined without a drill down.

4.9 *LCUG argues that fairness requires that the type 1 error be set larger than the conventional 2 1/2%. Again, this argument does not involve any assumptions regarding how the data actually behave.*

We agree that the issue is one of defining "fairness." We also agree that the setting of a significance level does not involve assumptions regarding the behavior of the data.

This issue of fairness, however, is not necessarily easy to resolve. The U.S. Supreme Court in *Castenada v. Partida*, 430 U.S. 482, 97 S.Ct. 1272 (1977) and *Hazelwood School District v. U.S.*, 433 U.S. 299, 97 S.Ct. 2736 (1977) adopted the rule that disparities should exceed 2 to 3 standard deviations in disparate impact cases. We have adopted "2" here -- the most common standard in general use.

4.10 *The final component of the FCC and LCUG approach concerns how a chosen level of type 1 error is to be achieved, by setting the critical value for the test. It is here that the form of the data does make a difference.*

We agree. In fact, as we stated, any testing should include acknowledgement of dependencies, if they exist across observations.

4.11 *To find the appropriate critical value, we must be able to derive the null-hypothesis distribution of the chosen test statistic - that is, the probability distribution of the values the statistic would take if the CLEC observations were in fact drawn from the same population as the BST ones. The BST analysts point out, correctly, that this distribution depends on the shape of the BST population; if this is Normal or close to Normal, then the textbook derivation applies and we can look up the critical values in published tables.*

We agree, except that an examination of the data shows that the BST data is far from Normal. However, by carefully assigning wire centers to replicates, Normal distribution theory can still be used on a test statistic whose variance estimate is based on the replicates.

4.12 *But if we do not have a Normal population, the textbook derivation does not apply. However, in the present case we do not need to make assumptions - we have data!*

The comment about not needing assumptions confuses us. It is true, of course, that when there is a large amount of data weaker assumptions may be possible. Our approach was a case of this. In particular, we checked the data, noted a wire center dependency, and used this knowledge to construct a test based on replication – a test with a minimal number of assumptions.

4.13 *For each data series, the BST analysts had access to large samples of BST data, and it would be completely straightforward to use the computer to draw many pseudo-random CLEC samples from these and so to derive the required distribution of the FCC or LCUG statistic.*

We agree that computer resampling techniques can be employed on this problem. That is, in fact our approach. Such techniques are, however, not necessarily "completely straightforward," especially if there are dependencies inherent in the data.

The replication method we have proposed does deal well with the dependencies we found. It relies on the computer to recalculate the same statistic for each replicate. Additional resampling and then averaging the results is promising. This is certainly in the spirit of Dr. Mallows' suggestion, and we intend to try more.

4.14 *Another requirement of the FCC and LCUG approaches is that the samples be independent. In Figure 10 and many subsequent Figures the BST analysts present evidence that there are differences among the wire-centers; for some wire centers, the provisioning interval tends to be large for both BST and the CLECs; for other wire centers, it is smaller.*

This is very important to recognize. We do not believe that deeper levels of disaggregation will eliminate this problem.

4.15 *This effect can easily be allowed for by relying on within-wire-center comparisons; this is what the BST analysts do, since they work with differences between BST and CLEC means within each wire center. The FCC and LCUG approaches can also handle this difficulty; we simply replace the overall variances by pooled within-wire-center variances. This is a completely standard form of adjustment.*

This may be true, but we do not believe that the within wire center variances are easy to compute. Remember, the BST and CLEC samples within a wire center are correlated. Thus, any calculation of a "pooled within-wire center" variance must include calculation of covariance terms. These may be very hard to analytically determine.

The alternative we have presented is a computer intensive technique that captures both within-wire center and between-wire center variation. Therefore, the testing can be done at a higher level of aggregation than the wire center.

4.16 *The effect of confounding variables, such as those the BST analysts discuss on pages B-5 and B-6, can also be allowed for in the FCC/LCUG approach. The BST team adjusts the data by using the weighted average D-hat in equation (3) (page B-7). This quantity could be used as the numerator of an FCC/LCUG statistic by matching it with a variance estimate computed from within-class variances.*

This is not the original form of the LCUG/Pooled approach that we had read about. We chose to use an approach that we have some expertise in applying.

4.17 *I therefore reject the conclusion of the BST analysts that the FCC and LCUG procedures have to rely on unwarranted assumptions. Once we have data, we do not need assumptions.*

We used an approach that we felt fit with the data that we had. By Dr. Mallows own admission above, the LCUG measure needs to be modified to handle the dependencies in the data. Our approach does this.

Statement No. 5

5.1 *The BST analysts use a Replicate Variance Estimation method to provide a scale on which to compare differences between BST and CLEC means. For each wire center, they compute the difference between the CLEC mean and the adjusted BST mean; they combine these into an overall estimate D-hat (equation (3), page B-7) using weights that correspond to the numbers of CLEC observations in each [difference].*

Actually, we compute the difference between means for each type of order, at each time, within each wire center. This is Cochran's method for dealing with observational data, and it provides an unbiased estimate of the difference between the means of BST and the aggregate CLECs.

5.2 *They then use the individual differences in equation (5) (page B-8) to get an estimate v_1 of the variance of the equally-weighted average of the differences, which they call d-double-bar.*

This is a common device used in replication. If in each replicate we have the same number of CLEC records, then the estimator is linear and $\hat{D} = \overline{\overline{d}}$. The assignment of wire centers to replicates is random, so if the sample size for the CLEC orders is large, we would expect that the estimator would be reasonably close to linear. If it is not close to linear, we can employ additional resampling techniques to correct this.

5.3 *However, since they want to use D-hat rather than d-double-bar as their overall estimate, they propose to replace v_1 by the estimate v_2 in equation (6).*

We chose the estimator recommended by Wolter[2], v_2. This was done with no further discussion because in the data we analyzed, there was no noticeable difference in the two estimates.

5.4 *This adjustment is in the wrong direction, and may be large. The effect is to favor BST by deflating the BST statistic.*

The following table gives the ratios of the standard error using v_2 to the standard error using v_1, for the estimated difference over all cases.

[2] Wolter, K. (1985) *Introduction to Variance Estimation*, Springer-Verlag, New York.

Measure - Month	$\sqrt{v_2 / v_1}$
OCI - August	1.0016
OCI - Sept	1.0057
OCI - October	1.0003
MAD - Aug	1.0052
MAD - Sept	1.0004

This shows that, for the data analyzed, that the "adjustment in the wrong direction" is not large.

5.5 *The estimate D-hat is more precise than is d-double-bar, and has a smaller variance; but the estimate v_2 is larger than v_1.*

We will make the suggested change by using $\hat{D} / \sqrt{v_1}$ as the basis for the test statistic rather than $\hat{D} / \sqrt{v_2}$.

5.6 *I cannot judge how big the effect is, this depends on how variable the CLEC sample sizes are, but I would not be surprised to find that the BST statistic has been deflated by a factor of 2.*

As we have shown in No. 5.4 above (and **Technical Appendix A.3**), the effect is not large. The reduction in the test statistic comes from taking wire center dependency into account when calculating the standard error of the difference of sample means.

Statement No. 6

6.1 *On page 15 the BST analysts say that "the BST analysis is designed to account for ... different standard deviations between BellSouth data and the CLECs". The BST analysis does not *account* for differing standard deviations.*

This is correct. In general, the BST method, as originally proposed, is not sensitive to the situation where the CLEC standard deviation is larger than the BST standard deviation; at the workshop we agreed to modify it. The details of this modification are presented in **Technical Appendix A.2**. Our new results provide a test that is equivalent to the LCUG test when the data are independent.

Incidentally, for the data that were analyzed, this adjustment would have made most of the tests **less significant** since the CLEC variance was generally smaller than the BST variance.

6.2 *Suppose for example that within each wire center, the BST and CLEC populations both have the same mean, but that the CLEC observations are more dispersed than the BST ones. See Illustration B, on page 7. Then parity of service is being violated within each wire center.*

While this hypothetical case is possible, the data we have looked at suggest that the opposite is true: the BST data are more dispersed than the CLEC data.

6.3 *The effect would be very hard to see in Figure 10; the CLEC means would be a little more dispersed than the BST means.*

This is true. We need to provide more diagnostics in order to check for this situation.

6.4 *The BST analysis, which uses only the differences between the BST and CLEC means, is completely insensitive to such differences. It would completely fail to detect such a violation of parity.*

We do not believe that such violations are present in the data we have analyzed. The modification that we propose for the test (see **Technical Appendix A.2**) will make it more sensitive to certain differences in BST and CLEC variances. Since the replicate method captures total variation in the data, a test of the hypothetical situation described would detect the significant difference in BST/CLEC performance.

6.5 *The BST analysts have not given us any information on the relative scales of the BST and CLEC variation within wire centers.*

We agree. We need to find ways to easily convey this type of information while respecting security concerns. The interpretation of such data, it might be noted, may be particularly challenging to interpret given that the wire centers are not identically distributed.

Statement No. 7

7.1 *On page B-3 the BST analysts assert that the "correct" test when the BST and CLEC variances are different is based on the statistic t' that they present at the top of the second column. The test based on t' is a test of the hypothesis that the means are equal, allowing the variances to be different. But this is not the appropriate null hypothesis. The t' test is not a test of the hypothesis that the BST and CLEC populations are the same.*

We agree on this point. But also add that the LCUG test is also not a test of the hypothesis that the BST and CLEC distributions are the same. The LCUG test is simply a test of the differences in means that has been modified to be sensitive to

certain situations where variances differ. We are modifying the BST test to have similar sensitivity (see **Technical Appendix A.2**).

Statement No. 8

8.1 *The simulation results that are reported on pages J-3-6 assume very large serial correlations - much larger than those found in Appendix G for the Average OSS differences.*

The correlations in the Interim Report's Appendix J are modeling dependencies between tests, not the serial or auto- correlation of a measure over time (which is what is looked at in Appendix G for the Average OSS Response Interval). As stated in Appendix J, the correlation structure was chosen because it has a uniform mix of correlation levels between parity measures.

8.2 *The Bonferroni method described on page J-6 assumes the worst possible correlation structure - in fact it allows for the possibility that the individual tests are perfectly correlated, they all pass or fail together.*

We agree that the Bonferroni method is conservative. That is why we do not recommend using it for more than 10 tests. It should be noted, however, that the procedure suggested by AT&T for 5 tests is approximately the same as the Bonferroni method.

8.2 *Empirical study is needed to check the degree to which the various tests are actually correlated.*

We agree. We need to study the correlation between measures. At this point in time we have only examined three measures from different SQM categories. By the time of the February Workshop we will have analyzed at least six (6) performance measures. Providing correlations across measures is planned.

8.3 *Regarding page J-3, the fact that the number of service requests varies comparatively smoothly for both BST and the CLECs does not imply that the FCC/LCUG statistics are correlated. We would need to look at the series of differences between BST and CLECs; this could easily resemble Figure 1 on page G-5, showing very little serial correlation.*

We agree that more study is needed to determine the autocorrelation of an individual test statistic from month to month. The last paragraph in Appendix J states this. While such an examination could easily show very little autocorrelation, it could also easily show that there is significant autocorrelation over time.

8.4 *The simulations on pages J-7-9 show that even for the extremely skew population in Figure 4 (I presume "Figure 1" on page J-7 is a misprint), the distribution of the LCUG statistic is close to standard normal except in the extreme tails.*

This was the point of the simulation, and it is one of the reasons we would not recommend using the Bonferroni method on more than 10 tests.